PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open. The paper was subsequently accepted for publication at BMJ Open.

(This paper received three reviews from its previous journal but only two reviewers agreed to published their review.)

ARTICLE DETAILS

TITLE (PROVISIONAL)	First-Incidence, Age of Onset Outcomes and Risk Factors of Onset of DSM-5 Oppositional Defiant Disorder: A cohort study of Spanish children from ages 3 to 9
AUTHORS	Ezpeleta, L; Navarro, J. Blas; de la Osa, Nuria; Penelo, Eva; Domènech, Josep Maria

VERSION 1 – REVIEW

REVIEWER	weidong ji Shanghai Changning Mental Health Center
REVIEW RETURNED	20-Mar-2018

GENERAL COMMENTS	The ideas in the paper are interesting and the double cohort design has some potential in applications of ODD. However, the statistical method and methodology need a big improvement, although the results are very helpful for the public health. Therefore, I do not recomend its acceptance as a full paper at this stage. Some suggestions may help the authors: 1. In the abstract, the conclusions should be rewritten for explaining what the result indicated. "Preventive interventions starting at preschool age are recommended" or "targeted and indicated interventions should be implemented to lessen the developmental difficulties and school and family burdens that cause ODD", you can not draw the conclusion from your results. 2. DSM-IV or DSM-5? The length of the follow-up period (7 years) was mentioned. Author should clarify the timeline, when was the diagnostic criteria changed? Why? Any difference? 3. Only nine out of 41 references article are within 3 years (2016-2018), innovation of the research is not well persuasive as a scientific report. 4. The worldwide-pooled prevalence of mental disorders was 13.4% (CI 95% 11.3-15.9). -- Annual Research Review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. Journal of Child Psychology and Psychiatry. But your team result showed"Up to 9 years old the cumulative risk of new cases of ODD was 21.9%". The pooled prevalence is 3.6% up to age 18 in child and adolescent, but the cumulative risk of ODD make me hard to understand and doubt whether there are overlaps or not. Further, "Cumulative risk was computed by the product-limit estimation using the weighted annual risk". Rate can not be cumulated, the statistical errors

	should be recount with proper range of age. 5. How were this large sample size study randomized? 2,283 children randomly selected should include a brief procedure in your method, if too long, please annotate in review contents. 6. In the “Patient and Public Involvement statement”, what’s the purpose of “Teachers received a 15 hours course about How to manage disruptive behavior disorder at the school-room at the beginning of different school levels (preschool -age 3-, elementary -ages 6 and 9)” but researchers did not arrange this for parents. Please describe the details and why. 7. In the table 2 and 3, the number of children decreased with stage of age for leaving the study except at age 8, which were more than at age 7, is that associated with the new increases? Please clarify it. 8. In the table 4,“Children with onset at 3-5 years old scored higher on all the scales scores of parent’s SDQ”, however, I can not find it even higher than that of 6-9 years old. 9. OR has been used to calculate in the table 4 while HR in Supplementary Table 1. Using one of them will make your paper more readable. 10. In the discussion, the risk of presenting ODD was highlighted, I still did not see psychiatrists or pediatricians who confirmed the diagnosis. The paper provided many scales for helping recognizing the characteristics of ODD, but as we know, scales merely play a role in auxiliary diagnosis. If the clinical diagnosis had not been explained clearly, even 7 years follow-up hardworking would have been less meaningful to readers. Neither did the random selection. Although the authors told readers the diagnostic information was based on the parents and teachers, a worldwide golden standard of DSM-5 can only be made through a doctor's diagnosis.

REVIEWER	Tamara Pringsheim University of Calgary
REVIEW RETURNED	11-Jul-2018

GENERAL COMMENTS	This research paper describes an epidemiological study of the incidence of ODD in children using a population based study design. This is an important study, due to the disability associated with this diagnosis. The paper is well written. The main issue that requires further clarification in both the methods and the results is their use of an enriched sample - 417 of the 622 children in their final sample screened positive for behavioural problems during the initial phase of the study. This would therefore inflate the risk of ODD diagnosis, and indeed the cumulative risk of ODD diagnosis in this sample is high, higher than has been reported elsewhere. What proportion of the 1,341 children in the first phase of sampling screened positive? Was the proportion of screen positives in the final phase higher than in the first phase? The authors state in their tables that their estimates are weighted by screen positive and screen negative membership- can this procedure of weighting be described? Can the authors provide the number of cases in the screen positive group and the screen negative group?

VERSION 1 – AUTHOR RESPONSE

Answers to Reviewer: 1

Please leave your comments for the authors below The ideas in the paper are interesting and the double cohort design has some potential in applications of ODD. However, the statistical method and methodology need a big improvement, although the results are very helpful for the public health. Therefore, I do not recomend its acceptance as a full paper at this stage. Some suggestions may help the authors:

1. **In the abstract, the conclusions should be rewritten for explaining what the result indicated. "Preventive interventions starting at preschool age are recommended" or "targeted and indicated interventions should be implemented to lessen the developmental difficulties and school and family burdens that cause ODD", you can not draw the conclusion from your results**. We have modified the conclusions and now the abstract says:

> **Conclusions** The risk of new cases of ODD in the general population at preschool age and during childhood is high. Preschool age is a target period for preventive interventions. Identified risk factors are objectives for targeted and indicated interventions.

2. **DSM-IV or DSM-5? The length of the follow-up period (7 years) was mentioned. Author should clarify the timeline, when was the diagnostic criteria changed? Why? Any difference?**

The reviewer is right that the DSM5 was published in 2013. The study started in 2019-10. As we use a diagnostic interview we had the information to write the algorithms in DSM5. As it was a longitudinal study, we needed to use the same diagnostic criteria for the sake of comparability. Therefore, all the diagnostic information was translated into DSM5 algorithms.
We indicate this when we describe the diagnostic interview (page 7, paragraph 2):

> The DICA-PPC(Ezpeleta, de la Osa, Granero, Doménech, & Reich, 2011) is a computerised semi-structured interview which generates diagnoses through algorithms following DSM-5 (American Psychiatric Association, 2013).

3. **Only nine out of 41 references article are within 3 years (2016-2018), innovation of the research is not well persuasive as a scientific report.**

We have included two new references of 2018: (Boekamp et al., 2018) and (La Maison et al., 2018).

4. **The worldwide-pooled prevalence of mental disorders was 13.4% (CI 95% 11.3-15.9). -- Annual Research Review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. Journal of Child Psychology and Psychiatry. But your team result showed "Up to 9 years old the cumulative risk of new cases of ODD was 21.9%". The pooled prevalence is 3.6% up to age 18 in child and adolescent, but the cumulative risk of ODD make me hard to understand and doubt whether there are overlaps or not. Further, "Cumulative risk was computed by the product-limit estimation using the weighted annual risk". Rate can not be cumulated, the statistical errors should be recount with proper range of age**.

There are not overlaps between the values of prevalence and cumulative risk because the risk is calculated considering new cases (incident cases) diagnosed with ODD in children without previous ODD (see Table 3). Risk is not a rate (Rothman, Greenland, & Lash, 2008). The cumulative risk it is computed through cumulative failure survival R(t)=1-S(t) estimated using the methods of actuarial survival analysis.

5. **How were this large sample size study randomized? 2,283 children randomly selected should include a brief procedure in your method, if too long, please annotate in review contents.**

We have included this description in page 6, Participants section:

> Children of each classroom were alphabetically numbered, and did not contain the name of the child nor the school. Those scoring under the cut-off were randomly permutated through the computer program SPSS (generating a random uniform number) and the first 30% was selected.

We include in the text Figure 1 showing the numbers of sampling selection and also the description of the randomization under Participants section.

6. **In the "Patient and Public Involvement statement", what's the purpose of "Teachers received a 15 hours course about How to manage disruptive behavior disorder at the school-room at the beginning of different school levels (preschool -age 3-, elementary -ages 6 and 9)" but researchers did not arrange this for parents. Please describe the details and why.**

In the "Patient and Public Involvement statement" it is necessary to indicate how the patients and the public were involved in the research. We understand that the teachers are individuals directly involved with the patients (the children). A way of involving the teachers in the research was through this 15 hours course. We did not arrange this course for parents because we considered that it was not appropriated in the context. Instead, for parents we considered other ways of implications more beneficial for them such as informing them individually and giving advice.

7. **In the table 2 and 3, the number of children decreased with stage of age for leaving the study except at age 8, which were more than at age 7, is that associated with the new increases? Please clarify it.**

The number of children usually decreased with stage of age because of attrition. At age 8 there was an increment instead a decrement due to two causes. Firstly, a child may not participate in the study for a given year but return to the study the following year (all the initial sample is invited every year to participate in the follow-up of that year). Second, if attrition is higher in a screening group that in the other (as happened at age 8), the weighting procedure could increase the total sample at that follow-up with respect to the previous.

8. **In the table 4,"Children with onset at 3-5 years old scored higher on all the scales scores of parent's SDQ", however, I can not find it even higher than that of 6-9 years old.**

The sentence mentioned by the reviewer is incomplete and the reviewer does not mention the comparison group. What the text says is (page 11, outcomes section):

> Children with onset at 3-5 years old scored higher on all the scales scores of parent's SDQ, higher conduct problems according to teachers, worse functioning and higher comorbidity **in comparison to children without ODD**.

According to statistical analyses, and as is shown in table 4 the sentence is correct.

However, we have included this sentence in the section Outcomes of age of onset of ODD
to clarify the results of the last column of Table 4:

> There were no differences in any SDQ score between preschooler and late ODD onset.

9. **OR has been used to calculate in the table 4 while HR in Supplementary Table 1. Using one of them will make your paper more readable.**

As stated in the statistical analysis section, the last rows of Table 4 show the results of logistic regressions models comparing the odds of having a DSM-5 diagnose between the three groups (no ODD, onset at 3-5 or at 6-9 years old). Therefore, the odds ratio (OR) is the appropriate statistic. Although it is possible to obtain an indirect estimation of the relative risk through the estimation of a Poisson regression model, we decided not to do it because of the potential bias.
Supplementary Table 1 shows the results of Cox regression modelling time to ODD diagnose as the outcome and different risk factor as predictors. There, the hazard risk (HR) is the appropriate statistic.

10. **In the discussion, the risk of presenting ODD was highlighted, I still did not see psychiatrists or pediatricians who confirmed the diagnosis. The paper provided many scales for helping recognizing the characteristics of ODD, but as we know, scales merely play a role in auxiliary diagnosis. If the clinical diagnosis had not been explained clearly, even 7 years follow-up hardworking would have been less meaningful to readers. Neither did the random selection. Although the authors told readers the diagnostic information was based on the parents and teachers, a worldwide golden standard of DSM-5 can only be made through a doctor's diagnosis.**

Clinical psychologists are qualified for making diagnosis of psychological problems. Structured and semi-structured diagnostic interviews administered by trained psychologists is an admitted and recommended method for making DSM and ICD diagnosis in epidemiological and research in general and also clinical work (Angold, Costello, & Egger, 2007; Angold et al., 2012). This is the way in which epidemiological studies with children and with adults use to work. The diagnoses in our study were carried out through a semi-structured diagnostic interview *Diagnostic Interview of Children and Adolescents for Parents of Preschool Children (DICA-PPC),* which generates diagnoses through algorithms following DSM-5 (American Psychiatric Association, 2013). Every year this interview was administered to parents and from parents information were derived the diagnosis of ODD, ADHD, major depression, any anxiety disorders (separation, generalized, social anxiety or specific phobias) at each age from 3 to 9 years old. These procedures are "standard" procedures in clinical psychology and in psychiatry.

Answers to Reviewer: 2

Please leave your comments for the authors below This research paper describes an epidemiological study of the incidence of ODD in children using a population based study design. This is an important study, due to the disability associated with this diagnosis. The paper is well written.

The main issue that requires further clarification in both the methods and the results is their use of an enriched sample - 417 of the 622 children in their final sample screened positive for behavioural problems during the initial phase of the study. This would therefore inflate the risk of ODD diagnosis, and indeed the cumulative risk of ODD diagnosis in this sample is high, higher than has been reported elsewhere. What proportion of the 1,341 children in the first phase of sampling screened positive? Was the proportion of screen positives in the final phase higher than in the first phase? The authors state in their tables that their estimates are weighted by screen positive and screen negative membership- can this procedure of weighting be described? Can the authors provide the number of cases in the screen positive group and the screen negative group?

We have includes Figure 1 that depicts the design of the study and the numbers participating in each point of the study. This, probably, will clarify the questions of the reviewer. The worry about ODD inflation is not justified because, precisely, what the weighting makes is to try to reproduce the original population from which the children come. The design of the study was a double phase design, very used in epidemiological studies, in which all the children upper the cut-off were selected and a proportion of children under the cut-off. Afterwards all the analyses are weighted to reproduce the original population distribution. The cut-off was based on conduct problems scores.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Association.

Angold, A., Costello, A. J., & Egger, H. L. (2007). Structured interviewing. In A. Martin & F. R. Volkmar (Eds.), *Lewi's child and adolescent psychiatry* (pp. 344-357). Philadelphia, PA: Lippincott-Williams & Wilkins.

Angold, A., Erkanli, A., Copeland, W., Goodman, R., Fisher, P. W., & Costello, E. J. (2012). Psychiatric diagnostic interviews for children and adolescents: A comparative study. *Journal of the American Academy of Child and Adolescent Psychiatry, 51*(5), 506-517.

Boekamp, J. R., Liu, R. T., Martin, S. E., Mernick, L. R., DeMarco, M., & Spirito, A. (2018). Predictors of partial hospital readmission for young children with oppositional defiant disorder. *Child Psychiatry and Human Development, 91*, 21-33. doi: 10.1007/s10578-017-0770-8

Ezpeleta, L., de la Osa, N., Granero, R., Doménech, J. M., & Reich, W. (2011). The Diagnostic Interview for Children and Adolescents for Parents of Preschool and Young Children: Psychometric Properties in the general Population. *Psychiatry Research, 190*, 137-144. doi: 10.1016/j.psychres.2011.04.034

La Maison, C., Munhoz, T. N., I.S., S., Anselmi, L., Barros, F. C., & Matijasevich, A. (2018). Prevalence and risk factors of psychiatric disorders in early adolescence: 2004 Pelotas (Brazil) birth cohort. *Social Psychiatry and Psychiatric Epidemiology, 53*, 685–697. doi: 10.1007/s00127-018-1516-z

VERSION 2 – REVIEW

REVIEWER	Peter Watson MRC Cognition and Brain Sciences Unit University of Cambridge UK
REVIEW RETURNED	06-Dec-2018

GENERAL COMMENTS	First-Incidence, Age of Onset Outcomes and Risk Factors of Onset of DSM-5 Oppositional Defiant Disorder: A cohort study if Spanish Children from ages 3 to 9. bmjopen-2018-022493.R1 Mainly a few queries about assumptions involving missingness and the Cox Model. Page 9, line 56 to Page 10, line 3 and Table 4 on Page 28. You are assuming in deleting missing values that these are occurring completely at random. Did you check this assumption of missingness by seeing if the missingness was related to any variables which might suggest an imputation of missing values? I notice in Table 4 on page 28 there appear to be larger quantities of missing values for the DSM-5 (lines 34-37) with between 25 and 47 missing (column 2). Is there a reason for these larger amounts of missingness? Page 11, line 5. Values of Harrell's C of at least 0.70 are regarded as good yet in Table 1 on page 31 only two of the Cox Models have values of Harrell's C greater than 0.70. Do the other models with poor fits therefore predict poorly the incidence of ODD and, if so, are these models trustworthy? Page 10, lines 51-53 and page 11 lines 1-3. The proportional hazards assumption is tested for and in cases where it is not met it states that separate estimates are made for different years (presumably of onset?). I don't see any further reference to the testing of this assumption in the results for the Cox Model (page 12, lines 27-49). Did all the models, therefore, satisfy the proportional hazards assumption and, if not, I wondered if there would be sufficient numbers of ODD cases to enable a powerful enough analysis fitting different Cox models to each year separately. I notice, for example, in Table 3 on page 27 that there are not very many (10-13) cases of ODD occurring in later years of onset (6-9).

	Page 10, lines 3-18 and Page 27 Table 3. I would explicitly say in the text and table that the Kaplan-Meier product-limit estimate was used to compute the cumulative risk. Page 14, line 31. The goal is to assess the predictive power of factors with regard to first onset of ODD. Is it possible, however, for a person to intermittently have ODD so that once they have ODD it doesn't mean they always will have ODD so that they can "come out of it" and become normal again? If ODD is intermittent does this have any rammifications for the use of the Cox Model? Does the Cox Model, for example, assume that once you have ODD you have it thereafter? Is it possible that a child could have already had ODD but this has not been recorded so that at the time of assessment they were incorrectly regarded as not having had a first incidence of ODD?

REVIEWER	Thomas Olino Temple University, USA
REVIEW RETURNED	18-Dec-2018

GENERAL COMMENTS	This is a manuscript examining emergence of ODD diagnoses across early to middle childhood and predictors of those disorders. The work relies on a large sample of youth with a number of longitudinal assessments. This permits identifying new cases over time. I was asked to provide a statistical review of the work. In this capacity, I see the authors being quite clear in several of their responses. The motivation for their survival and logistic regression models makes sense. The area of the manuscript that I found to be less clear, methodologically, is on the sample weighting. Could more be said about the implementation? Based on the multiple staging of sampling, it was not clear how the weights would be derived. The results are interpreted by the authors as indicating a continued need for early identification and intervention. Based on these data, the authors are in good position to describe the timeline for when *most* cases of ODD may emerge. I think that a plot of lifetime and incident ODD may be useful to illustrate this take home message more clearly than the data presented in Tables 2 & 3. The results for associations between early onset and current functioning are confounded by duration of illness and current status. Are there means of strengthening the analyses to identify whether these alternatives are driving these results? Are inter-rater reliability data available for the diagnostic interviews?

VERSION 2 – AUTHOR RESPONSE

Answers to Reviewer: 3

Reviewer Name: Peter Watson

Mainly a few queries about assumptions involving missingness and the Cox Model.

Page 9, line 56 to Page 10, line 3 and Table 4 on Page 28. You are assuming in deleting missing values that these are occurring completely at random. Did you check this assumption of

missingness by seeing if the missingness was related to any variables which might suggest an imputation of missing values?

Response: A global analysis looking for randomness of attrition during the 7 years of follow-ups with respect to sex, type of school (state or private) and socioeconomic status was already done as indicated in Participants section.

> No differences in sex (χ^2 =0.07; *p* =.793) or type of school (χ^2 =0.72; *p* =.396) were found on comparing completers and drop-outs during the seven years of annual follow-ups. However, the SES of those leaving the study until age 9 was lower (χ^2 =20.89; *p* <.001).

Additionally, we have verified that attrition in each annual follow-up was not related to the screen group defined in the first sampling stage. The next sentence has been added in the Participants section:

> The percentage of drop-outs at annual follow-up from ages 4 to 9 was similar in the two screen groups (χ^2 = 0.72, *p* = .798 at age 4; χ^2 = 0.31, *p* = .575 at age 5; χ^2 = 1.36, *p* = .244 at age 6; χ^2 = 0.02, *p* = .877 at age 7; χ^2 = 0.49 and *p* = .484 at age 8; χ^2 = 0.20 and *p* = .652 at age 9).

The analysis mentioned above showed evidence that missingness was randomly. As Little and Rubin (1987) stated, checking if missing values are completely at random is not possible usually because it implies knowing the unknown values. However, because at the first follow-up we have complete data for the whole initial sample (N=622), we have compared the outcome scores at age 3 between cases and drop-outs at age 9. For 10 out of the 16 outcomes shown in Table 4 there were no significant differences. For 6 outcomes the score at age 3 was higher for drop-outs than for completers at age 9. The results of this new analysis could be considered a strict approach for checking if missing data are completely at random. The next paragraph has been added in the Participants section:

> Finally, to assess randomness of attrition the outcome scores at age 3 between cases and drop-outs at age 9 were compared. For 6 out of the 16 outcomes scores at age 3 were higher for drop-outs than for completers at age 9.

And also a sentence has been added as one limitation:

> A second limitation refers to the non-randomness of attrition in 6 out of the 16 outcomes analyzed as risk factors of first ODD diagnose. However, as shown in several populations, attrition is associated with adverse psychosocial variables and high levels of psychological distress (Fischer, Dornelas, & Goethe, 2001; Granero, Ezpeleta, & Doménech, 2007)

I notice in Table 4 on page 28 there appear to be larger quantities of missing values for the DSM-5 (lines 34-37) with between 25 and 47 missing (column 2). Is there a reason for these larger amounts of missingness?

Response: The higher level of missing values in DSM-5 variables in Table 4 is due to the interview format for obtaining these variables. Interview requires a longer time for responding than questionnaires. Some families accepted to answer the questionnaires but they did not have the time to dedicate to a longer diagnostic interview.

Page 11, line 5. Values of Harrell's C of at least 0.70 are regarded as good yet in Table 1 on page 31 only two of the Cox Models have values of Harrell's C greater than 0.70. Do the other models with poor fits therefore predict poorly the incidence of ODD and, if so, are these models trustworthy?

Response: As indicated in the objective of the manuscript, the results presented in Table 1 online were related to "… test if previously reported risk factors associated with ODD are prospectively risk factors of incident cases at these developmental stages". In this sense, although it is true that models with Harrell's C lower than .70 make poor predictions of the incident ODD cases, we consider those models worthy because most of their predictors are statistically significant, showing a relevant effect.

To clarify the predictive analysis the next paragraph has been added at the end of the Results section:

> The capability to predict new ODD first-incident cases from the subsets of risk factors was low in general. Only the first model with "being an ODD substreshold" as predictor, and the second model with "ODD Irritability and Headstrong" as predictors showed Harrell's *C* ≥ .70.

Also, in the Discussion section we indicate:

> Predictive capability assessed by Harrell's C was generally low to moderate, **indicating that to predict first-incident ODD cases other predictors are needed in addition to the clinical risk factor considered.**

Page 10, lines 51-53 and page 11 lines 1-3. The proportional hazards assumption is tested for and in cases where it is not met it states that separate estimates are made for different years (presumably of onset?). I don't see any further reference to the testing of this assumption in the results for the Cox Model (page 12, lines 27-49). Did all the models, therefore, satisfy the proportional hazards assumption and, if not, I wondered if there would be sufficient numbers of ODD cases to enable a powerful enough analysis fitting different Cox models to each year separately. I notice, for example, in Table 3 on page 27 that there are not very many (10-13) cases of ODD occurring in later years of onset (6-9).

Response: The proportional hazards assumption refers to the equality of effect of predictors over time. When it was not meet, the HR estimates were made separately for each year of follow-up. To clarify this aspect the sentence in Statistical Analysis section has been rewritten:

> In the presence of significant interaction, the hazard ratio (HR) for the involved predictor was obtained separately for each year of follow-up, corresponding to ages 3 to 8.

All the models except the one with the two scales from Children's Behavior Questionnaire satisfied the proportional hazards assumption. This was indicated in the explanation accompanying Table 1 online:

> As the effect of CBQ negative affect did not meet the proportional hazard assumption, a HR was obtained for each year.

Consequently, Table 1 online shows independent estimates for the CBQ negative affect effect for the ages 3 to 8.

As the reviewer indicates, the number of ODD cases in the last 4 years of the study was low. This has been included as a limitation:

> Finally, the fact that increasing the age of the children, the number of incident cases diminished, limited the statistical power.

Page 10, lines 3-18 and Page 27 Table 3. I would explicitly say in the text and table that the Kaplan-Meier product-limit estimate was used to compute the cumulative risk.

Response: The detail about Kaplan-Meier has been added in the text and in the footnote of Table 3:

> … cumulative risk was computed by **the Kaplan-Meier** product-limit estimation (Kaplan & Meier, 1958) using the weighted annual risk.

Page 14, line 31. The goal is to assess the predictive power of factors with regard to first onset of ODD. Is it possible, however, for a person to intermittently have ODD so that once they have ODD it doesn't mean they always will have ODD so that they can "come out of it" and become normal again? If ODD is intermittent does this have any rammifications for the use of the Cox Model? Does the Cox Model, for example, assume that once you have ODD you have it thereafter? Is it possible that a child could have already had ODD but this has not been recorded so that at the time of assessment they were incorrectly regarded as not having had a first incidence of ODD?

Response: As the title indicates, our work analyzes the first-incidence of an ODD diagnosis. Therefore, all the analysis related to incidence excluded second, third or posterior ODD diagnose. This was already indicated in a footnote of Table 3:

> [a]Incident cases (after excluding children with previous diagnoses of ODD)

Also, at the end of the Participant section:

> Decrements in sample size at successive follow-ups were either due to attrition or to the exclusion of children who had already presented a first ODD diagnosis.

And finally, in the Statistical Analysis section. This sentence includes now a specification about the restriction of only studying first ODD diagnose:

> The incidence proportion was calculated for 1-year time period beginning at 4 years old by dividing the number of **new cases of first ODD diagnose** (incident cases) by the number of children at risk, i.e. the number of cases at the beginning of the period excluding those who had previous diagnoses of ODD.

Responses to Reviewer: 4

Reviewer Name: Thomas Olino

This is a manuscript examining emergence of ODD diagnoses across early to middle childhood and predictors of those disorders. The work relies on a large sample of youth with a number of longitudinal assessments. This permits identifying new cases over time.

I was asked to provide a statistical review of the work. In this capacity, I see the authors being quite clear in several of their responses. The motivation for their survival and logistic regression models makes sense. The area of the manuscript that I found to be less clear, methodologically, is on the sample weighting. Could more be said about the implementation? Based on the multiple staging of sampling, it was not clear how the weights would be derived.

Response: The final sample (*N* = 622) was formed by adding all the 417 children from the positive screen-group and a random sample of 205 children from the negative screen-group. Consequently, the final sample over-represented the positive screen-group with the aim of incrementing the sample prevalence of conduct problems. To compensate this over-estimation the statistical analysis was weighted by assigning each child a value that was inverse to the probability of random selection in the second phase of sampling.

To clarify this aspect, the explanation of the second sampling stage in the Participants section has been rewritten:

> The final sample for the follow-ups (second-phase) included 622 children (Figure 1) comprising all the children from the screen-positive group whose families accepted to participate (*N* = 417; 49.4% boys) and a random sample from the screen-negative group (N = 205; 51.2% boys). To select participants from screen-negative group children of each classroom were alphabetically numbered without including the name of the child nor the school. Then they were randomly permutated using SPSS random number generator, and the first 30% was selected.

Also, the explanation of the statistical analysis has been modified to better relate the weighting procedure:

> Since all the data were collected using a double-phase screening design, all analyses were weighted by assigning each child a value that was inverse to the probability of random selection in the second phase of sampling.

The results are interpreted by the authors as indicating a continued need for early identification and intervention. Based on these data, the authors are in good position to describe the timeline for when *most* cases of ODD may emerge. I think that a plot of lifetime and incident ODD may be useful to illustrate this take home message more clearly than the data presented in Tables 2 & 3.

Response: Following the recommendation of the reviewer, a new Figure 2 showing prevalence of ODD and incidence of first ODD diagnose has been added.

The results for associations between early onset and current functioning are confounded by duration of illness and current status. Are there means of strengthening the analyses to identify whether these alternatives are driving these results?

Response: Following the recommendation of the reviewer, the results of Table 4 have been re-analyzed adjusting the models by current ODD diagnosis and by number of years with an ODD diagnosis, additionally to having or not ODD treatment (already included in the previous version). Some results have changed; therefore the discussion has been modified.

Are inter-rater reliability data available for the diagnostic interviews?

Response: In the training process of the interview (see (Ezpeleta, de la Osa, Granero, Domènech, & Reich, 2011) we require as criterion for being ready for the field to obtain a mean agreement with an expert kappa ≥ .80. Inter-rater was tested in a pilot with 13 interviews when the psychometric properties of the interview were analyzed (Ezpeleta, de la Osa, & Doménech, 2014). The details of the pilot study were:

Inter-interviewer agreement was studied in a pilot study with 13 interviews of children from public pediatric primary care, whose families accepted to participate.
Mean age of the children was 5.54 years (SD = 0.97; range 4-7 years-old) and 7 (53.85%) were boys. All the interviewers were trained following the procedure described later. For each interview the kappa coefficient (Cohen, 1960) between the interviewer and each of the observers (raters) in the main diagnoses of interest of the interview was calculated. The interviews were recorded and, afterwards, were rated by 5 team interviewers resulting in 65 observations. All the interviewers took the role of "interviewer" and "raters". Kappa coefficients of the rating ranged from .83 to 1 (mean kappa .92; range between .83 and 1) (being .84 for ODD) indicating a good to excellent agreement between interviewers.

We have included this sentence in the Measures section when we describe the interview:

> Inter-interviewer agreement in the diagnoses ranged from Kappa coefficients from .83 to 1 (mean kappa .92; being .84 for ODD), indicating a good to excellent agreement between interviewers.

Ezpeleta, L., de la Osa, N., & Doménech, J. M. (2014). Prevalence of DSM-IV disorders, comorbidity and impairment in 3-year-old Spanish preschoolers. *Social Psychiatry and Psychiatric Epidemiology, 49*(1), 145-155. doi: 10.1007/s00127-013-0683-1

Ezpeleta, L., de la Osa, N., Granero, R., Domènech, J. M., & Reich, W. (2011). The Diagnostic Interview of Children and Adolescents for Parents of Preschool and Young Children: psychometric properties in the general population. *Psychiatry Research, 190*, 137-144. doi: 10.1016/j.psychres.2011.04.034

Fischer, E. H., Dornelas, E. A., & Goethe, J. W. (2001). Characteristics of people lost to attrition in psychiatric follow-up studies. *Journal of Nervous and Mental Disease, 189*(1), 49-55. doi: 10.1097/00005053-200101000-00009

Granero, R., Ezpeleta, L., & Doménech, J. M. (2007). Features associated with non-participation and abandonment in mental health epidemiological designs of socially-at-risk children and adolescents. *Social Psychiatry and Psychiatric Epidemiology, 42*, 251-258.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457-481.

Little, R.J.A. & Rubin, D.B. (1987). *Statistical analysis with missing data*. New York: Wiley

VERSION 3 – REVIEW

REVIEWER	Thomas Olino Temple University
REVIEW RETURNED	08-Feb-2019

GENERAL COMMENTS	This is a revision examining the emergence of ODD across ages 3 to 9. In their revision, the authors have been responsive. The additional figure displaying the incidence and prevalence of ODD is helpful for communicating much information efficiently. I have no further queries.